D318–D319 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 25 September 2009*
doi:10.1093/nar/gkp786

PDBselect 1992–2009 and PDBfilter-select

Sven Griep and Uwe Hobohm*

University of Applied Sciences Giessen, Bioinformatics, D-35390 Giessen, Germany

Received August 1, 2009; Accepted September 4, 2009

ABSTRACT

PDBselect (http://bioinfo.tg.fh-giessen.de/pdbselect/) is a list of representative protein chains with low mutual sequence identity selected from the protein data bank (PDB) to enable unbiased statistics. The list increased from 155 chains in 1992 to more than 4500 chains in 2009. PDBfilter-select is an online service to generate user-defined selections.

In 1992, bioinformatics was a tiny discipline *in statu nascendi*. Neither existed bioinformatics as a university discipline nor were the skills shaped, which are required to work at the edge between biology and computer science. Chris Sander's group at EMBL-Heidelberg was one of the few bioinformatics groups in Europe and consisted mainly of biologists and a few physicists. Programming abilities of those days bioinformaticians were self-aquired, the time slice needed for debugging within a programming project usually was considerable and the programming tools were in its infancy. Programs were edited in emacs and debugged from the UNIX shell, Fortran programs were not able to read input from the terminal without writing a linefeed character and jumping the waiting cursor to an empty line, sequence searches in SwissProt or EMBL were done using Smith-and-Waterman or Pearson's Fasta program, the Internet was yet to come. Our statistics on protein structures were done on a Sun Sparc-Station I cycling around 20 MHz and benchmarking at ~1.5 MFlops—a contemporary Core-2-Duo processor is running with 2.5 GHz and benchmarks at ~600 MFlops. We wanted to save time and restrict statistical analysis of protein structures to a representative subset of structures.

PDBselect started, when the realm of protein chains with known 3D structure was around 700 or 0.6% of the July 2009 count, resulting in a representative list of 155 protein chains with mutual sequence similarity of <30%(in subsequent releases we used a threshold of 25%). We decided to use protein chains as entity since the best granularity to cut protein structures into domains was unclear and chain breaks resulting from less well-resolved regions of the protein structure appeared to be natural divisors separating more densely packed building blocks. To generate the representative list of protein chains, an all-versus-all sequence comparison was implemented. The distance between two protein sequences is calculated by applying the HSSP function (1), later refined by Abagyan and Batalov based on a larger data set(2). When two protein chains score related by the function, the one with lower quality is removed, to end up with a representative list of high-quality structures. Quality was defined as 'resolution [in Angstrom] plus [R-factor (in percent)/20]', with NMR structures allocated an arbitrary (low) quality. I (U.H.) remember us arguing whether 20 is the appropriate value, yet, this constant has not much influence on the size of the final list. One of the successors of PDBselect, ASTRAL (3), uses a similar measure supplemented by stereochemical checks from PROCHECK (4) and WHAT_CHECK (5).

The introduction of the HSSP function caused some irritation. Meticulous users found that some protein pairs within the 25%-representative list had sequence identities above 25%. But this is a consequence of the function, which uses both alignment length and sequence identity to score alignments. For long alignments, a sequence identity as low as 22% may still indicate homology, while for short alignments a higher sequence identity above 45% may be required to infer homology. Of course, all this is exercise in 1D certainly missing many 3D homologies, but it is still too expensive to perform an exhaustive 3D-all-on-all comparison on a regular basis.

Algorithm and parameters were exchanged or adjusted over time. While initially, we used Smith-and-Waterman(6) for the alignments with quadratic RAM space requirements, later we switched to the faster Huang–Miller algorithm (7) with linear RAM space demand. Chain selection is now done by sorting chains on 'quality' with best chains on top of the list, and then looping down: take the first chain, eliminate all subsequent homologs, take the next non-eliminated chain, eliminate all homologs, a.s.o. which turns out to be faster than the prior greedy algorithm (8,9). The PAM-matrix was replaced by Blosum-65, gap elongation penalty was changed from 4 to 1(10) resulting in more restrictive, shorter lists. Chain data like resolution, R-factor, chain amino acid sequence are, for easier retrieval, accessed not from PDB-files directly but from the derived flat file PDBfinder(11).

PDBfilter-select (http://bioinfo.tg.fh-giessen.de/pdbselect/pdbfilter-select/pdbfilter-select.pl) was implemented to

*To whom correspondence should be addressed. Tel: 0641 3092580; Fax: 0641 3092549; Email: uwe.hobohm@tg.fh-giessen.de

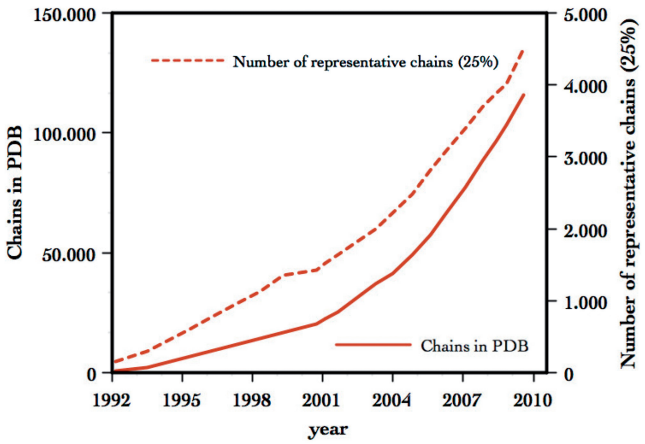

Figure 1. Increase in number of PDB chains (solid line, left *Y*-axis) and representative chains (dashed line, right *Y*-axis) 1992–2009.

meet the wish of many users for custom-made selections via a web interface. In a first filtering step, a pre-selection of chains is generated, where chains can be enriched based on the length, method (X-ray, NMR, fiber, neutron, other), resolution, R-factor, number of residues in helix or beta conformation. Out of this pre-selection, a PDBselect run generates a list of non-homologous chains. Due to CPU restrictions, the starting list is limited to a size of 3000 chains; however, larger pre-selections can be generated upon request.

While it took a couple of days in 1992 to calculate the PDBselect list on a Sun-Sparc station, it takes 2 days in 2009 on an Apple Mac-Pro tower. Between 1992 and today, PDBselect—which in the beginning was implemented as a little helper serving us other purposes—has acquired hundreds of citations and a long list of subscribers. The ratio between number of all chains and number of chains in the representative list decreased from 1/5 in 1992 over 1/14 in 2000 to 1/26 in 2009, expressing the decreasing likelihood to find novel structures. Still, saturation is not in sight, the 18-year plot of PDBselect shows no tendency of plateauing (Figure 1), indicating that we are not yet close to complete coverage of protein structures by the PDB. Thus, the vastness of protein space will unfold for a while. PDBselect shall follow (http://bioinfo.tg.fh-giessen.de/pdbselect).

FUNDING

The Open Access publication charges for this paper has been waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
2. Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.
3. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Aacids Res.*, **32**, D189–192.
4. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
5. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
6. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
7. Huang,X. and Miller,W. (1991) A time efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, **12**, 337–357.
8. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
9. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
10. Zachariah,M.A., Crooks,G.E., Holbrook,S.R. and Brenner,S.E. (2005) A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins*, **58**, 329–338.
11. Hooft,R.W., Sander,C., Scharf,M. and Vriend,G. (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. *Comput. Appl. Biosci.*, **12**, 525–529.